DOCUMENT RESUME

ED 350 310 TM 019 005

AUTHOR Marsh, Herbert W.; Bailey, Michael
TITLE Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis.
PUB DATE 24 Jul 91
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Feedback; Foreign Countries; Formative Evaluation; Generalization; Graduate Study; Higher Education; *Professors; *Profiles; Rating Scales; Scores; *Student Evaluation of Teacher Performance; Summative Evaluation; Teacher Characteristics; *Teacher Effectiveness
IDENTIFIERS *Students Evaluations of Educational Quality

ABSTRACT

The Students' Evaluations of Educational Quality (SEEQ) instrument measures distinct dimensions of teaching effectiveness. Previous research has shown that ratings of a given instructor tend to generalize over different course offerings, but research has not examined whether each instructor had a generalizable, distinguishable profile of scale scores (e.g., high on enthusiasm but low on organization). This study considered 3,079 sets of class-average responses for 123 instructors collected over a 13-year period. The profile of 9 SEEQ scales (learning/value, instructor enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments/readings, and workload/difficulty) for each instructor was distinct from those of other instructors, was consistent across course offerings over the 13-year period, and generalized across undergraduate and graduate level courses. The results have important implications for the use of profiles as formative feedback, summative evaluations, and further research on students' evaluations of teaching. Four graphs illustrate the discussion, and there is one table of analysis results. An appendix contains factor analysis results for 24,158 sets of data. Nineteen references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED350310

Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HERBERT W. MARSH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Herbert W. Marsh

University of Western Sydney, Macarthur

and

Michael Bailey

University of Sydney

24 July, 1991

Running Head: Profile Analysis

TM019005

BEST COPY AVAILABLE

ABSTRACT

The Students' Evaluations of Educational Quality (SEEQ) instrument measures distinct dimensions of teaching effectiveness. Previous research has shown that ratings of a given instructor generalize over different course offerings, but research has not examined whether each instructor has a generalizable, distinguishable profile of scale scores (e.g., high on enthusiasm but low on organization). Considered here were 3079 sets of class-average responses for 123 instructors collected over a 13-year period. The profile of 9 SEEQ scales (e.g., Enthusiasm, Organization, Group Interaction) for each instructor was distinct from those of other instructors, was consistent across course offerings over the 13-year period, and generalized across undergraduate and graduate level courses. These results have important implications for the use of profiles as formative feedback, summative evaluations, and further research on students' evaluations of teaching.

The literature on students' evaluations of teaching effectiveness (SETE) consists of 1000s of studies and dates back to the 1920s and earlier (Marsh, 1987). In reviews of this research, Marsh (1982, 1984, 1987; Marsh & Dunkin, in press) concluded that SETEs are: (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases to the ratings; and (f) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, by administrators for use in personnel decisions, and by researchers. Based on his review, Marsh (1987) argued that SETEs are probably "the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research" (p. 369).

SETEs, like the teaching that they represent, are a multidimensional construct (e.g., an instructor may be organized but lack enthusiasm). This contention is supported by common-sense and a considerable body of empirical research (Marsh, 1987). The SETE literature contains several examples of well constructed instruments with clearly defined factor structures that provide measures of distinct SETE scales. In his review, Marsh (1987) noted that the systematic approach used in the development of these instruments and the similarity in the factors identified by each supports their construct validity. The Students' Evaluation of Educational Quality (SEEQ) instrument that is the focus of the present investigation appears to measure the most broadly representative set of scales and to have the strongest factor analytic support of these instruments (Marsh, 1982, 1983, 1984, 1987). Marsh and Hocevar (1991) factor analyzed SEEQ responses for a total group of 24,158 courses and for each of 21 different subgroups selected to differ in terms of instructor rank, course level (undergraduate and graduate), and academic discipline. The set of 9 factors that SEEQ is designed to measure (see Appendix 1) was identified in all 22 factor analyses, and factor scores based on the total group were highly correlated with factor scores based on separate analyses of each of the 21 subgroups (mn $r > .99$). The results demonstrated that SEEQ measures a broadly representative set of evaluation factors and supported the generalizability of the SEEQ factor structure.

A logical extension of the multidimensionality of SETEs is that a given instructor has a distinct profile of SETE ratings (e.g., high in organization but low in enthusiasm) that generalizes over time and across different courses. There is apparently no research that has examined this hypothesis and so the purpose of this investigation is to test it. Support for existence of a distinguishable profile that is specific to each instructor has important implications for understanding SETEs, for the use of SETEs as feedback, and for the relation of SETEs to other criteria such as student learning. The necessary starting point of such research is to determine whether instructors have distinguishable SETE profiles and this is the purpose of the present investigation.

Studies of the generalizability of SETEs have typically considered global ratings, total scores, or separate analyses of specific SETE dimensions. Researchers (e.g., Bausell & Bausell, 1979; Kulik & Kulik, 1974; Marsh, 1981; Gilmore, Kane, & Smith, 1981) have examined correlations between ratings of the same instructor in different offerings of the same course, the same instructor in different courses, and different instructors teaching the same course in an attempt to disentangle the relative influence of the course and the instructor. This research examined the generality of SETEs and the relative importance of the instructor who teaches the course and the course that is being taught. For example, for overall instructor and course ratings, the correlations between ratings of different instructors teaching the same course (one estimate of the course effect) were -.05 and -.01 respectively, whereas correlations between ratings for the same instructor in two different classes (.61 and .59) and in two different offerings of the same course (.72 and .71) were much larger (Marsh, 1981). These findings demonstrated that overall SETEs were generalizable and that SETEs were primarily a function of the instructor rather than the course.

In a recent extension of this research, Hanges, Schneider and Niles (1990) evaluated the stability of SETEs over 13 semesters. Although conceptually similar to earlier research aimed at distinguishing between course effects and teacher effects, Hanges et al. focused on an interactionist framework from social and personnel psychology. According to this framework, performance is a function of the person (the instructor) and the situation (the course being taught). In order to test this hypothesis, Hanges et al correlated: (a) ratings of the same instructor (averaged across all different courses) in 13 different semesters -- "person" correlations; (b) ratings of the same course (averaged across all instructors) in 13 different semesters -- "situation" correlations; and (c) ratings of the same instructor in the same course -- "person in situation" correlations. The findings were largely consistent with those reported by Marsh (1987) in that the "person in situation" correlations were modestly larger than "person" correlations, whereas "situation" correlations were substantially smaller. The authors concluded that "the weak stability coefficients for the situation analysis suggest that, at least for faculty behavior in the classroom, faculty attributes rather than attributes of the courses taught yield stability in ratings" (p. 665). However, by apparently averaging across all ratings of the same course in each semester to obtain a situation effect, the so-called situation effect actually confounded agreement in ratings of the same course taught by different instructors (an appropriate estimate of the situation effect) and agreement between ratings of same course taught by the same instructor (an estimate of the "person in situation" effect). Hence, even though their estimate of their situation effect was very small, apparently it was positively biased and an unbiased estimate would have been even smaller.

Discussing limitations in existing research and directions for future study, Marsh (1987) noted that the generality of mean ratings of separate SETE scales had been examined but not the generality of SETE profiles.

Whereas Hanges, et al. (1990) and Marsh (1981) evaluated different SETE scales, separate analyses were conducted on each scale. No attempt was made, for example, to determine whether the instructor who is enthusiastic but lacks organization for one course receives a consistent pattern of ratings in other courses. Marsh (1987) noted that more research was needed to determine the generalizability of profiles of scores as well as the generalizability of each score within the profile. He argued that the existence of reliable individual differences in profiles of SETE dimensions would have important theoretical and practical implications. If each instructor has a distinguishable profile that is distinct from the typical profiles of other instructors, then researchers must examine how entire profiles of scores are related to different validity criteria (e.g., student learning) as well as examining the relation of each score within the profile. We know of no research, however, that has pursued this important issue. Hence, the primary purpose of the present investigation is to determine whether or not there are individual differences in the SETE profiles that generalize across ratings of the same instructor collected over a 13-year period.

In profile analysis (Tabachnick & Fidell, 1989), it is important to distinguish between the level of a profile (e.g., whether the SETE ratings across all the scales are consistently high or low for a particular instructor) and the shape of a profile (e.g., whether each instructor has a distinguishable profile of scores). For present purposes this distinction reflects the difference between the generalizability of mean ratings (level) and the generalizability of profiles of ratings (shape). In the terminology of profile analyses, generalizability studies of SETEs have considered the generalizabiltiy of profile levels but have not considered the generalizability of profile shape. The purpose of the present investigation is to extend existing research by evaluating the generalizability of profile shapes.

Methods

Sample and Procedures.

During the 13-year period 1976-1988, almost one million SEEQ forms were administered in approximately 50,000 courses at a large private university in western United States. Although the use of the SEEQ form is voluntary, the university requires that each academic unit collect some form of SETEs for all courses. Staff are not considered for promotion unless SETEs are provided. Most of the academic units that do use SEEQ require that all their staff are evaluated in all courses. The evaluation forms are typically distributed to staff shortly before the end of each academic term, administered and collected by a student in the class or by a member of the administrative support staff according to printed instructions, and taken to a central office where they are processed. This program, the SEEQ instrument on which it is based, and research that led to its development are described by Marsh (1987).

SEEQ (Marsh, 1987) measures 9 evaluation factors (see Appendix 1 for the wording of SEEQ items) that have been supported by more than 40 factor analyses (e.g., Marsh, 1983, 1984, 1987; Marsh & Hocevar, 1991; also see

Appendix 1). In the present investigation, a set of 9 factor scores based on the factor analysis summarized in Appendix 1 was used to represent the profile of SEEQ scores for each set of class-average ratings considered here. For present purposes, instructors were selected who had: (a) been evaluated at least once during each of 10 different years between 1976 and 1988 and (b) been evaluated in at least two graduate level courses and two undergraduate level courses by 10 or more students. This selection process resulted in the selection of 123 different instructors who had been evaluated in 3,079 classes -- an average of about 25 classes per instructor.

Statistical Analysis.

Profile analyses (e.g., SPSS, 1988; Stevens, 1986; Tabachnick & Fidell, 1989; Timm, 1975) are traditionally conducted using either a univariate repeated-measures analysis or a conceptually similar multivariate analysis of variance (MANOVA) [1]. Both the repeated-measures and multivariate approaches to profile analysis address three questions (e.g., Tabachnick & Fidell, 1989):

1. Level. Are there systematic instructor differences on the total SETE scores (i.e., the average of all nine SEEQ scores)? In the traditional repeated measures analysis this tests whether the total scores that the same instructor receives are consistent across different course offerings and vary systematically from the total scores received by other instructors. This is the type of question typically pursued in previous studies of the generalizability of SETE responses. Although this is not the primary focus of the present investigation, it is a very important question.

2. Parallelism. Are the profiles of SEEQ scales for different instructors parallel or do the profiles vary systematically for different instructors? In the traditional repeated measures analysis this refers to whether differences in SEEQ scales (the repeated measures variable) interact with the instructor variable. The effect tests whether there are systematic individual differences in the profiles of SEEQ scores. Hence, the significance and size of this effect are the most important results of this study.

3. Flatness. This question asks whether there are systematic differences in the 9 SEEQ scores averaged across all instructors. This question is not of concern here and, as noted by Tabachnick and Fidell (1989), the question is typically irrelevant unless the assumption of parallelism is accepted. As suggested by Stevens (1986) and by Tabachnick and Fidell (1989), all SEEQ scores were standardized (mean = 0, SD = 1) to eliminate this effect. Because the SEEQ scores were based on factor scores, the scores already were standardized. Even if unstandardized scale scores had been used, however, comparisons of the (raw) mean scores for different scales would not be justified. Thus, for example, the comparison of any two scales for a given teacher would reflect some unknown combination of teaching characteristics measured by each scale and the particular wording of the items constructed to reflect each scale. By first standardizing all scales across a broad normative sample, the comparison reflects how a given teacher compares with the total sample on each of the two scales.

For purposes of the present investigation, analyses were conducted using both repeated measures and MANOVA approaches. Because conclusions based on the two approaches are essentially the same, the focus will be on the repeated measures approach although both sets of results are presented. In these analyses, each "case" was one of the 3,079 sets of 9 SEEQ scales representing a particular course offering taught by a particular instructor. The design factors in the analyses were the instructor (the 123 different instructors), the course level (undergraduate or graduate), and the set of 9 SEEQ scales. Because instructor and course level were crossed (i.e., each instructor had sets of ratings at both graduate and undergraduate levels) these were "between" factors, whereas the set of 9 SEEQ scores was a repeated measures factor (i.e., each case contained ratings for all 9 SEEQ scales). The between factors assessed the influence of the instructor, the course level, and their interaction on the total SEEQ score (i.e., the total of the SEEQ scales). A large instructor effect, for example, would indicate that the SEEQ total score is largely determined by the instructor who is being evaluated, as has been found in previous research (see earlier discussion of profile "level"). The main effect of the within factor was necessarily zero since all the SEEQ scales were standardized to have the same mean and standard deviation (see earlier discussion of profile "flatness"). The interaction of the within and between factors assessed whether the effect of between factors varied depending on the SEEQ scale. A large instructor by scales interaction, for example, would indicate that there are systematic differences among instructors in the profiles of SEEQ scales (see earlier discussion of profile "parallelism"). Because of well-known problems in assessing statistical significance using the univariate repeated measures analysis, the Greenhouse-Geisser Epsilon, the Huynh-Feldt, and the maximally conservative Lower-bound epsilon (SPSS, 1988; Tabachnick & Fidell, 1989) were used. However, because of the very large sample size, all effects were statistically significant ($p < .001$) using even the maximally conservative lower-bound test, so that alternative tests of statistical significance are not substantively important. For this reason, the focus of discussion is not on statistical significance per se, but rather effect sizes (Eta squared and omega squared). The corresponding MANOVA was similar except that statistical significance of the corresponding effects were tested with Wilk's Lambda (SPSS, 1988).

Results and Discussion

Before presenting the detailed statistical analyses, it is useful examine the profiles for a few teachers from our study. The profiles of four teachers (Figure 1) illustrate the "level" and "shape" comparisons that are the focus of our profile analyses. Each profile is the average ratings across all sets of ratings for the same teacher collected during the 13-year period. Instructors 1 and 2 have generally higher ratings than instructors 3 and 4, demonstrating the effects of level. The effect of shape can be seen by comparing the Enthusiasm and Organization scores for the different teachers. Instructors 1 and 3 have consistently higher ratings for Organization than Enthusiasm, whereas instructors 2 and 4 have consistently higher ratings for Enthusiasm than for Organization. These comparisons demonstrate the separate effects

of level and shape that are evaluated more systematically in the univariate repeated measures analysis and the corresponding MANOVA summarized in Table 1.

Insert Table 1 About Here

Instructor and Course Level Differences on the SEEQ Total Scores.

The main effects of instructors, course level, and their interaction tests the generality of ratings received by the same instructor over different course offerings (i.e., profile levels). This analysis is conceptually equivalent to doing an analysis on the total score across the nine SEEQ factors. Because this portion of the analysis is not based on multiple dependent measures, the univariate and MANOVA approaches are equivalent. Although the effect sizes vary somewhat depending on whether eta ^{2} or omega ^{2} is used, the instructor effect is large and much larger than the effects of either the course level or the course level by instructor interaction.

The very large effect of the instructor is the most important of these results. The eta ^{2} of .371 is equivalent to an average correlation of .61 between ratings of the same instructor on different course offerings. This result is consistent with Marsh's (1981) finding that the average correlation was .71 for two offerings of the same course and .52 for offerings of two different courses. The findings here, perhaps, are more impressive because the generality is across different ratings of the same instructor over a 13-year period of time. Both studies are consistent in providing support for the mean generalizability of SETEs.

Although not a focus of the present investigation, the main effect of course level is consistent with other research (e.g., Marsh, 1984; 1987) showing that graduate level courses are evaluated somewhat more favorably than undergraduate level courses. The present results are, perhaps, more compelling in that the difference in course level is based on ratings by the same instructors in graduate and undergraduate level courses. The course level by instructor interaction effect, however, has apparently not been examined previously (but see Murray, Rushton, & Paunonen (1990)). This effect suggests that some instructors get systematically higher ratings in graduate level courses whereas other instructors get systematically larger ratings in undergraduate courses. Whereas these effects are conceptually important, it should be emphasized that they are much smaller than the effect of the instructor.

Instructor and Course Level Differences on SEEQ Profiles.

Analyses in this section examine the extent to which SEEQ profiles are parallel across all the instructors, and across the graduate and undergraduate courses. In the repeated measures analysis this is determined by the extent to which instructor differences vary depending on the specific SETE scale (i.e., the instructor by scales interaction in Table 1). Although the results vary somewhat depending on the effect size that is considered, the

effect of the instructor by scale interaction is large and much larger than the course level by scale interaction or the course level by instructor by scale interaction. The very large effect of the instructor demonstrates that most of the systematic variance in profiles, and nearly 50% of the total variation, is due to the specific instructor. Particularly since the different SEEQ dimensions are moderately correlated (see Appendix 1), these substantial effect sizes demonstrate that the profiles associated with different instructors are distinct. These results indicate that there is a very good consistency in the profiles obtained by the same instructor on different course offerings over a 13-year period. This finding is the most important result of the present investigation and the major focus of the study.

The course level by scale interaction, though statistically significant, accounts for very little variance. This means that, averaged across all instructors, the profiles of ratings are similar in graduate and undergraduate classes. Whereas the earlier results showed that total scores were somewhat higher in graduate level courses, these results indicate that the effects are similar across the different SEEQ scales. The instructor by course level by scale interaction, however, indicates that there are some differences in the profiles of the same instructor in graduate and undergraduate courses. These effects, however, are much smaller than the effect of instructor.

Summary and Implications

The most important result of this investigation is that there are large and systematic differences in the SEEQ profiles obtained by different instructors. The SEEQ profile for each instructor generalizes across sets of ratings obtained by the same instructor over a 13-year period and across undergraduate and graduate level courses. The results also showed that there are large and systematic differences in the SEEQ total scores obtained by different instructors. The SEEQ total score for each instructor generalizes across sets of ratings of the same instructor over a 13-year period and across undergraduate and graduate level courses. Although these results are superficially similar, they are quite separate and statistically independent. The results for the profiles refer to the shape of the profile for each instructor after removing the effect of level (i.e., the total score). The results for the total scores refer to level and is independent of the shape of the profile. The large effect of instructors on the total scores supports the generality of the SETEs across multiple course offerings, but does not say anything about the multidimensionality of the ratings. The large effect of instructors in the analysis of profiles, however, provides very strong support for the multidimensionality of SETEs. In summary, the findings suggest that instructors appear to have distinct profiles of strengths and weaknesses that are highly generalizable, and that students are apparently able to discriminate their instructors' strengths and weaknesses, at least when ratings are aggregated over many students. The present investigation is apparently the first test of the generality of SETE profiles and in this respect provides important new evidence in support of the multidimensionality of SETEs.

The presentation of the results emphasized the separate effects of the between- and within-sections of the

analysis. Also, the effect sizes for total scores and for profiles are each based on portions of the variance in the two different sections of the analysis. However, when the number of scores in the profile is as large as in the present investigation, the comparison of the two results is somewhat misleading. The total sums of squares for the within effects is substantially larger than for the between effects. Hence, the consistency in instructor profiles explains much more variance than the consistency of instructor total scores -- even more than is implied by a comparison of the respective effect sizes. As is the case with any interaction effects, these results indicate that differences between instructors vary substantially depending upon the specific SEEQ scale that is being considered so that total score differences must be interpreted cautiously. This conclusion, of course, is consistent with an emphasis on the multidimensionality of SETEs and the importance on SETE profiles.

The findings of the present investigation open up new, and largely unexplored areas for further research on the implications of SETE profiles. Previous validity research has focused on relations of external validity criteria to either overall SETEs or, sometimes, specific scales of SETEs. Because there are reliable individual differences in SETE profiles that are independent of SETE levels, our results imply that researchers should also consider how different profiles are related to external criteria of effective teaching. For example, it may be that student learning is maximized when both Enthusiasm and Organization are high, whereas being high on just one or the other is insufficient. Similarly, the demonstration of systematic, reliable individual differences in SETE profiles supports the use of profiles in formative feedback and summative evaluations and has important implications for the study of teaching and teaching styles.

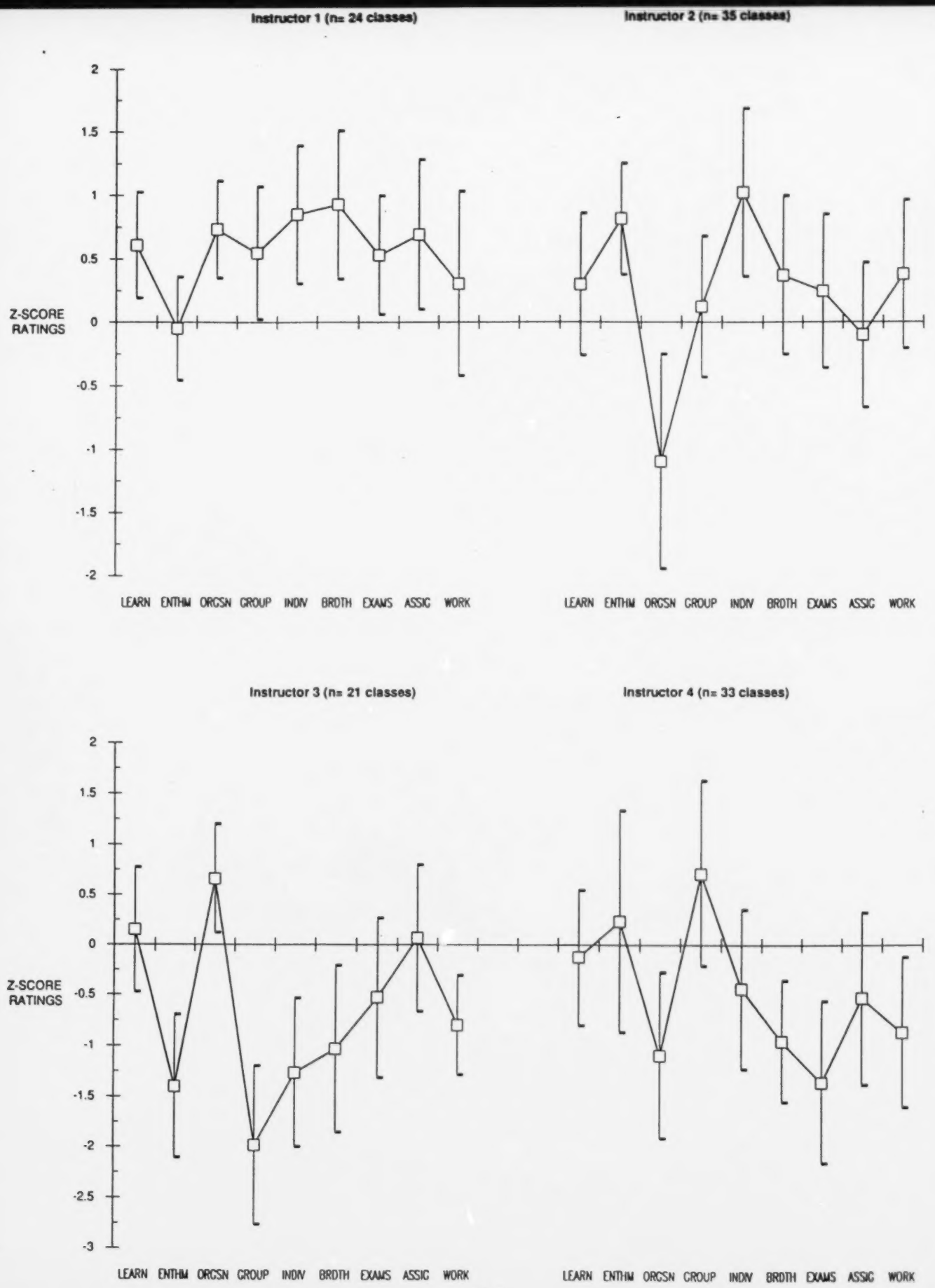
Although not a major focus of the present investigation, the results also have important implications for the broader study of personnel and organizational psychology. There is an on-going concern about the reliability and stability of performance ratings, and their situational specificity (e.g., Hanges, Schneider & Niles, 1990). The results of the present study, along with earlier research (Marsh, 1987), provide a clear example in which evaluations of performance are reliable, generalizable, reasonably stable over long periods of time, and reasonably independent of the specific situation (i.e., the particular course) in which the ratings are collected. More specifically, the results demonstrate an application of profile analysis that may have broad applicability in personnel psychology. In the same way that university teachers have distinct profiles on multiple dimensions of teaching effectiveness, individuals in other performance evaluation settings may have distinct profiles on multidimensional indicators of effectiveness that are reliable and generalizable. If individual differences in profiles do exist, then researchers should consider how different profiles -- as well differences in the specific dimensions that comprise the profiles -- are related to external criteria of effectiveness and to other outcome measures.

Footnotes

[1] As noted by an anonymous reviewer, this approach to profile analysis resembles -- at least conceptually -- the multivariate extension of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1981). Another anonymous reviewer noted some similarity between this type of analysis and confirmatory factor analysis approaches to multitrait-multimethod data (e.g., Boruch, 1970; Marsh, 1987; Marsh & Hocevar, 1984).

Figure Caption

Figure 1. Profiles of nine SEEQ scales for 4 instructors (see Appendix for more detail on the SEEQ scales). All scores were standardized (mean = 0, SD = 1) across all sets of ratings used in the study. Each profile represents the mean score for each SEEQ scale (the boxes), averaged across all the classes for that instructor. Thus, for example, all scores above the line representing a z-score of zero reflect ratings that are above average. Also presented for each scale, is the range of scores corresponding to the mean plus and minus one standard deviation (based on the set of ratings for the particular instructor for that particular scale).


Instructor 1 (n= 24 classes)
Instructor 2 (n= 35 classes)
Instructor 3 (n= 21 classes)
Instructor 4 (n= 33 classes)
Z-SCORE RATINGS
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
LEARN ENTHM ORGSN GROUP INDIV BRDTH EXAMS ASSIG WORK

REFERENCES

Bausell, R. B., & Bausell, C. R. (1979). Student ratings and various instructional variables form a within-instructor perspective. Research in Higher Education, 11, 167-177.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Wiley: New York.

Gilmore, G. M., Kane, M. T., & Smith, M. L. (1981). The generalizability of student ratings of instruction: Estimates of teacher and course components. Journal of Educational Measurement, 15, 1-13.

Hanges, P. J., Schneider, B., & Niles, K. (1990). Stability of performance: An interactionist perspective. Journal of Applied Psychology, 74, 658-667.

Kulik, J. A., & Kulik, C. (1974). Student ratings of instruction. Teaching of Psychology, 1, 51-57.

Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. Applied Psychological Measurement, 6, 47-60.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology, 52, 77-95.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research International Journal of Educational Research, 11, 253-388. (Whole Issue No. 3)

Marsh, H. W. (1989). Confirmatory factor anlaysis of multitrait-multimethod data: Many problems and a few solutions Applied Psychological Measurement, 13, 335-361. (Whole Issue No. 3)

Marsh, H. W., & Dunkin, M. J. (in press). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (ed.), Higher education: Handbook of theory and research. New York: Agathon.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. American Educational Research Journal, 21, 341-366.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. Teaching and Teacher Education: An International Journal of Research and Studies, 9-18.

Murray, H. G., Rushton, J. P. & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. Journal of Educational Psychology, 82, 250-261.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Erlbaum.

Tabachnick, B. G. & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). New York: Harper & Row.

Timm, N. H. (1975). Multivariate statistics with applications in education and psychology. Belmont, CA: Brooks/Cole.

Table 1

Univariate Repeated Measures and MANOVA Analyses of SEEQ Profiles

Repeated Measures Analysis					MANOVA			
			Effect Sizes[a]					
Source	SS	df	η^2 eta	ω^2 omega	Wilks' Lambda	Hypoth df	Error df	Mutivariate F
Between (Total Scores)								
Instructor (I)	3894	122	.371	.349				
Course Level (L)	166	1	.016	.016				
I x L	934	122	.089	.067				
Error Between	5493	2833						
Total Between	10488	3078						
Within (Profiles)								
Scales x I	8099	976	.470	.451	.0050	976	22594	23.60
Scales x L	108	8	.006	.006	.8923	8	2826	42.65
Scales x I x L	1195	976	.069	.050	.3403	976	22594	3.31
Error Within	7812	22664						
Total Within	17214	24632						

Note. The repeated measures and MANOVA approaches for the Total Scores (i.e., the Between Groups portion) are equivalent since there is only one dependent measure. For both approaches the 9 SEEQ scores were transformed into 8 difference scores between adjacent SEEQ factors -- the standard "repeated" transformation in SPSS (1988). Whereas the ordering of the SEEQ scores is arbitrary, the results in no way depend on the particular ordering used (see Tabachnick & Fidell, 1989, for further discussion).

[a] The Greenhouse-Geisser, Huyhn-Feldt, and Lower-bound Epsilons were 0.872, 0.950, and 0.125 respectively. Even when the Lower-bound Epsilon that is known to be maximally conservative was used , all tests of statistical significance were significant at $p < .001$.

Appendix 1

Factor Analysis Results For the Total Sample of 24,158 Sets of Class-average Responses: Factor Loadings and Factor Correlations. (Adapted with permission from Marsh & Hocevar, 1991).

SEEQ Scales and Items (paraphrased)	SEEQ Factors								
	Lrn	Enth	Orgn	Grp	Ind	Brd	Exam	Asgn	Work
Learning/Value									
Course challenging & stimulating	|.434|	.168	.103	.015	.014	.159	.099	.155	.291
Learned something valuable	|.607|	.083	.100	.026	.050	.103	.085	.147	.113
Increase subject interest	|.646|	.078	.034	.039	.058	.169	.074	.131	.020
Learned & understood subject matter	|.487|	.043	.176	.152	.045	.047	.112	.149	-.217
Overall Course Rating	|.410|	.211	.173	.041	.042	.085	.166	.175	.069
Instructor Enthusiasm									
Enthusiastic about teaching	.095	|.544|	.129	.072	.195	.115	.052	.069	.025
Dynamic and energetic	.064	|.714|	.094	.059	.085	.083	.069	.071	.042
Enhanced presentation with humor	.089	|.650|	-.023	.103	.078	.129	.090	.054	-.045
Teaching style held your interest	.137	|.581|	.187	.131	.026	.050	.110	.073	.017
Overall Instructor Rating	.172	|.392|	.245	.083	.141	.096	.140	.075	.039
Organization/Clarity									
Lecturer explanations clear	.146	.165	|.510|	.176	.060	.075	.079	.104	-.072
Materials well explained & prepared	.069	.087	|.677|	.060	.075	.073	.094	.118	.005
Course objectives stated & pursued	.128	.026	|.529|	.055	.070	.065	.175	.184	.024
Lectures facilitated taking notes	.031	.040	|.589|	-.093	.049	.175	.146	.044	.020
Group Interaction									
Encouraged class discussion	.058	.103	.011	|.769|	.070	.033	.067	.080	.002
Students shared knowledge/ideas	.066	.049	-.015	|.797|	.095	.093	.048	.073	-.029
Encouraged questions & gave answers	.059	.105	.167	|.583|	.151	.094	.100	.080	.001
Encouraged expression of ideas	.045	.069	.035	|.674|	.182	.110	.094	.070	-.013
Individual Rapport									
Friendly towards individual students	.051	.163	-.001	.176	|.612|	.063	.112	.057	-.038
Welcomed students seeking help/advice	.042	.059	.061	.078	|.786|	.036	.093	.059	-.007
Interested in individual students	.086	.140	.001	.137	|.647|	.057	.138	.059	.004
Accessible to individual students	-.014	-.028	.139	.037	|.636|	.099	.136	.104	.010
Breadth of coverage									
Contrasted various implications	.043	.037	.118	.059	.068	|.676|	.077	.109	.065
Gave background of ideas/concepts	.087	.085	.134	.020	.044	|.662|	.056	.122	.004
Gave different points of view	.035	.066	.086	.123	.101	|.636|	.097	.113	-.004
Discussed current developments	.207	.113	.018	.086	.039	|.562|	.084	.040	.000

Examinations/Grading									
Examination feedback valuable	.034	.039	.111	.047	.101	.028	|.670|	.088	.044
Evaluation methods fair/appropriate	.047	.044	.011	.043	.107	.078	|.749|	.099	-.033
Tested course content as emphasized	.063	.036	.129	.034	.064	.047	|.643|	.146	-.026
Assignments/Readings									
Readings/texts were valuable	-.008	-.004	.019	.022	.018	.053	.025	|.885|	-.003
They contributed to understanding	.127	.021	.036	.027	.039	.012	.140	|.716|	.072
Workload/Difficulty									
Course difficulty (easy-hard)	-.028	.030	.051	-.059	-.017	.096	.015	.018	|.861|
Course workload (light-heavy)	.100	-.054	.004	.085	-.001	.002	-.035	.038	|.907|
Course pace (slow-fast)	-.098	.101	.055	-.099	.005	-.001	.035	.040	|.689|
Hours per week outside of class	.148	-.044	-.085	.034	-.001	-.006	-.006	.042	|.798|

Factor Pattern Correlations	Lrn	Enth	Orgn	Grp	Ind	Brd	Exam	Asgn	Work
Learning/Value	1.000								
Instructor Enthusiasm	.434	1.000							
Organization/Clarity	.407	.427	1.000						
Group Interaction	.350	.364	.210	1.000					
Individual Rapport	.263	.400	.331	.455	1.000				
Breadth of Coverage	.449	.419	.454	.327	.352	1.000			
Examinations/Grading	.401	.392	.511	.315	.493	.403	1.000		
Assignments/Readings	.488	.319	.431	.312	.338	.418	.510	1.000	
Workload/Difficulty	.128	.076	.044	-.072	-.009	.106	.033	.154	1.000

Note. Target loadings, the factor loadings items designed to define each SKEQ factor, are presented in boxes. The statistical analyses were performed with the commercially available SPSS (1988) factor analysis procedure using principal axis factor, a Kaiser normalization, and a direct oblimin (oblique) rotation. Factor scores were generated by SPSS using the regression method for all 24,158 sets of class-average responses -- including the 3,079 sets of class-average responses used in the present investigation.